Using machine learning models to predict oxygen saturation following ventilator support adjustment in critically ill children: a single center pilot study.

Sam Ghazal[1], Michael Sauthier[2], David Brossier[2], Wassim Bouachir[3], Philippe Jouvet[2*] and Rita Noumeir[1]

[1] Department of health information analysis, École de Technologie Supérieure (ÉTS), Montreal, Quebec, Canada

[2] Department of Pediatrics, Sainte-Justine Hospital, Montreal, Quebec, Canada

[3] LICEF research center, TÉLUQ University, Montreal, Quebec, Canada

* Corresponding author:

philippe.jouvet@umontreal.ca (PJ)

Abstract

Clinicians' experts in mechanical ventilation are not continuously at each patient's bedside in an intensive care unit to adjust mechanical ventilation settings and to analyze the impact of ventilator settings adjustments on gas exchange. The development of clinical decision support systems analyzing patients' data in real time offers an opportunity to fill this gap. The objective of this study was to determine whether a machine learning predictive model could be trained on a set of clinical data and used to predict hemoglobin oxygen saturation 5 min after a ventilator setting change. Data of mechanically ventilated children admitted between May 2015 and April 2017 were included and extracted from a high-resolution research database. More than 7.10^5 rows of data were obtained from 610 patients, discretized into 3 class labels. Due to data imbalance, four different data balancing process were applied and two machine learning models (artificial neural network and Bootstrap aggregation of complex decision trees) were trained and tested on these four different balanced datasets. The best model predicted SpO_2 with accuracies of 76%, 62% and 96% for the SpO_2 class "< 84%", "85 to 91%" and "> 92%", respectively. This pilot study using machine learning predictive model resulted in an algorithm with good accuracy. To obtain a robust algorithm, more data are needed, suggesting the need of multicenter pediatric intensive care high resolution databases.

Introduction

In case of respiratory failure, mechanical ventilation supports the oxygen (O_2) diffusion into the lungs and the carbon dioxide (CO_2) body removal. As an expert in mechanical ventilation cannot reasonably be expected to be continuously present at the patient's bedside, specific medical devices aimed to help in ventilator settings adjustments may help to improve the quality of care. Such devices are developed using either algorithms based on respiratory physiology/medical knowledge that adapt ventilator settings in real time based on patients' characteristics but are not accurate enough to be used widely in clinical practice, especially in children [1, 2]; or physiologic models that simulate cardiorespiratory responses to mechanical ventilation settings modifications but none was validated for this indication [3]. The above-mentioned models all share the limitation of not being suited to learn from ever-growing sets of clinical research data, and potentially improve their performances. To overcome this drawback, another avenue is the development of algorithms using artificial Intelligence to provide caregivers with support in their decision-making tasks. In this study, we assessed machine learning methods to predict transcutaneous hemoglobin saturation oxygen (SpO_2) of mechanically ventilated children after a ventilator setting change using a high resolution research database.

Materials and Methods

This study was conducted at Sainte-Justine Hospital and included the data collected prospectively between May 2015 and April 2017 of all the children, age under 18 years old, admitted to the Pediatric Intensive Care Unit (PICU) who were mechanically ventilated with an endotracheal tube. Patients' data were excluded if the patient was hemodynamically unstable defined as 2 or more vasoactive drugs delivered at the same time (ie., epinephrine, norepinephrine, dopamine or vasopressin) or with an uncorrected cyanotic heart disease defined by no SpO_2 > 97% during all PICU stay. All the respiratory data from included patients were extracted from the PICU research database [4], after study approval by the ethical review board of Sainte-Justine hospital (number 2017 1480).

Data extraction

To determine the data that will be extracted for each child, an item generation was conducted by three physicians (PJ, MS, DB). The resulting items are presented in Fig 1 within their sources, means of extraction and a schematic of the main components of the study. The predictive SpO_2 value was the SpO_2 5 minutes after a change of a ventilator setting. The delay of 5 min corresponded to the shortest period of time to reach a steady state after modification of a ventilator setting [5].

Fig 1. Schematic description of the analysis process and items involved. EMR: electronic Medical Record, FiO_2: inspired fraction of Oxygen, Vt: tidal volume, PEEP: Positive end expiratory pressure, PS above PEEP: pressure support level Above PEEP, PC above PEEP: pressure control level above PEEP, MVe: expiratory minute volume, I:E Ratio: inspiratory time over expiratory time, Measured RR: respiratory rate measured by the ventilator, PIP: positive inspiratory pressure ie maximal pressure measured during inspiration. $_{5min}SpO_2$: SpO_2 observed 5 min after PEEP, FiO_2, tidal volume, PS above PEEP, PC above PEEP change, ML: machine learning, ANN: artificial neural network, BACDT: Bootstrap aggregation complex decision trees.

Data formatting

The data extracted from the research database needed: (1) to remove erroneous data due to disconnection of the patient from the ventilator or the monitor, or due to transient interventions such as suctioning; (2) to remove the rows at which no ventilator setting variables was modified; (3) to adapt data format for classifier training. The methodology to format the data is described in S1 file.

Data categorization

SpO_2 levels at 5min were classified into three categories (Table 1). The thresholds were selected according to clinical value: a $SpO_2 < 92\%$ is a target to increase oxygenation in mechanically ventilated children [6]. The critical level of 85% SpO_2 is used as an alarm of severe hypoxemia in intensive care [7].

Table 1: Definition of SpO_2 class labels specifications

SpO_2 classification	SpO_2 range (%)	Rows number (n)
1	< 84	17,112
2	85 to 91	29,869
3	92 to 100	729,746

Data balancing

The data analysis showed a severe imbalance with most SpO_2 at 5min above 92%. This is logical as caregivers want to maintain SpO_2 in normal range during child PICU stay. In such condition, the classifier learns the majority class label (class 3) (Table 1) but doesn't learn the minority class labels (class 1 and 2) [8]. The data balancing process aims to allow the classifier to learn from all class equally. The data balancing process used in this study included a combination of down-sampling and up-sampling techniques: to balance the three classes of the data involved, a down-sampling of the SpO_2 class 3 using TOMEK algorithm [9] and an over-sampling of SpO_2 class 1 and 2 using Synthetic Minority Oversampling Technique (SMOTE) [10] were performed.

The creation of synthetic data points by SMOTE can be formulated as follows:

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \tag{2}$$

In equation (2), x_{syn} represents the synthetic data point. The variables x_i and x_{knn} are respectively the original instance, and the nearest neighbor data point which is randomly picked among the k nearest neighbors.. The random number δ is generated in [0,1] to determine the position of the created synthetic data point along a straight line joining the original data point x_i and its chosen nearest neighbor x_{knn}.

To study which data balancing method provided the more accurate algorithm, four datasets were produced via four different balancing procedures, involving different combinations of data balancing techniques (Fig 2).

Fig 2. Descriptions of the four balancing procedures.

Predicted SpO_2 Classification

To identify the best machine learning classification method, we tested two classification models: artificial neural network and bagged complex decision trees, on the four balanced datasets.

Artificial Neural Network (ANN)

Once the data has been pre-processed, a machine learning predictive model was trained on a sub-set of labeled training data. The model is then used to predict the target variable values on a testing subset where the class labels are hidden. We used Artificial Neural Networks (ANN) to make predictions of the SpO_2 variable, based on the values of other variables of interest. Through the function approximation that the ANN performs, it is possible to make predictions of SpO_2 variable, based on the input data.

The ANN is learned from training data, using the backpropagation algorithm [11] and is tested on a test set made of the remaining rows of data to validate the generalization of the model. The learning algorithm runs through all the rows of data in the training data set and compares the predicted outputs with the target outputs found in the training data set. The weights are adjusted via supervised learning, in a manner to minimize the error of predicted SpO_2 vs target SpO_2. The process is repeated until the error is minimized.

The ANN classifier was implemented through cycles of forward propagation followed by backward propagation through the network's layers. The backpropagation algorithm is used for performance optimization.

For a given number of classes K > 2, the cross-entropy error can be formulated as shown in eq. 3, where $(\boldsymbol{W}_i)_i$ is the matrix of weights between the neuron layers, r_i is the target value. y_i is the value generated by the ANN, ie., its output.

$$E^t((\boldsymbol{W}_i)_i \,|\, x^t, r^t) = - \sum_i r_i^t \log y_i^t \tag{3}$$

The outputs of the ANN are:

$$y_i^t = \frac{\exp w_i^t x^i}{\sum_k \exp w_k^t x^t} \tag{4}$$

Using stochastic gradient-descent (SGD) for error minimization, the update rule for the ANN weights is:

$$\Delta w_{ij}^t = n\left(r_i^t - y_i^t\right)x_j^t \tag{5}$$

In equation 5, η is the learning rate, which, when SGD is used, decreases as the error is minimized. During ANN training, each observation, comprised of an input vector and a target output, is denoted ($\boldsymbol{x}^t, \boldsymbol{r}^t$), with $\boldsymbol{r}^t \in$ ("1", "2", "3"). The reason why the cross-entropy (eq. 3) is used instead of the Least Square Error (LSE) is to avoid long periods of training, due to the ANN going through stages of slow error reduction.

Bootstrap aggregation of complex decision trees

Bootstrap aggregating (acronym: bagging) was proposed by L Breiman in 1994 to improve classification by combining classifications of randomly generated training sets [12]. Bagging allows for the creation of an aggregated predictor via the use of multiple training sub-sets taken from the same training set. Let ($\boldsymbol{T}^i$) denote the replicate training sub-sets bootstrapped from the training set $\boldsymbol{T}$. These replicate sub-sets each contain N observations, drawn at random and with replacement from $\boldsymbol{T}$. For each of these sub-sets of N observations, a prediction model, or classifier, is created. The computational model we used for bagging was complex decision trees. This means that, for each bootstrapped sub-set of training data, a complex decision tree is trained and thus a classifier is created. If $i = 1, \ldots, n$, then n classifiers are created through the bagging process.

A decision tree is a flowchart computational model which can be used for both regression, as well as classification problems. Paths from the root of the tree to its various leaf nodes go through decision nodes in which decision rules are applied in a recursive manner, based on values of input variables. Each path represents an observation $(\boldsymbol{X}, y) = (x_1, x_2, x_3, \ldots, x_n, y)$, where the label assigned to the target y is given in the leaf node, at the end of the path, ie., classification [13].

In the aim of maximizing the model's generalization capability during the training process, the Bagged Complex Trees' performance is tested via *k*-fold cross-validation. A value $k = 10$, which is common practice, was used in this study. The training using *k*-fold cross-validation is carried out as described in Fig 3.

Fig 3. *k*-fold cross-validation

The mathworks Matlab R2016b Machine Learning toolbox was used for the creation of the ensemble of Bagged complex trees model.

Assessment of the performances of the classifiers

We evaluated the performances of the classifiers based on the metrics including testing confusion matrix, average accuracy, precision, recall and F score [14] with a $_{5min}SpO_2$ prediction expected above 0.9 for each class.

- Precision

$$Precision = \frac{\#\,True\ positives\ class\ i}{Total\ \#\,classifications\ for\ class\ i} \tag{6}$$

The *Precision* (eq. 6) is the ratio of all correct classifications for class *i* to all instances labeled as class label *i* by the model. In a non-normalized confusion matrix, this would mean dividing the number of instances classified in class label *i* by the total of instances in column *i*.

- Recall

$$Recall = \frac{\#\,True\ positives\ class\ i}{Total\ \#\,observations\ class\ i} \tag{7}$$

Recall is the ratio of the number of instances classified in class label i to the number of true class i labels. In a non-normalized matrix, this would require dividing the number of instances classified in class label i by the total of row i

- F-score

$$F - score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \tag{8}$$

The F-score provides a single measure of classification performance of the model used.

Results and discussion

We developed and assessed the performances of two machine learning classifiers on four different balanced datasets to predict SpO_2 at 5 min after a ventilator setting change (*ie* FiO_2, PEEP, Vt/Pressure), in 610 mechanically ventilated children. In Fig 4 and Table 2, we report the performances of these two classifiers. Using the classification performance metrics, the bagged trees classifier trained on dataset #3 (see Fig 2) has yielded the best classification performance on the test sets (Table 2). The confusion matrix of the whole bagged trees shows that SpO_2 at 5 min could correctly predict in **76%** of class “1” data, **62%** of class “2”, and **96%** of class “3” (Fig 4). This huge variation in classification performances of the three class labels can be explained by the large variation in the numbers of observations available for each of the class labels in the initial dataset that has limited the machine learning (Table 1).

Fig 4. Artificial neural network (ANN) and bootstrap aggregation of complex decision trees (BACDT) test confusion matrices. The darker colors represent higher levels of accuracy. A:

balanced dataset 1, B: balanced dataset 2, C: balanced dataset 3, D: balanced dataset 4 (see Fig 2).

Table 2. Performance of artificial neural networks (ANN) and bootstrap aggregation of complex decision trees (BACDT) classifiers for SpO_2 prediction at 5 min following a ventilator setting change. Avg/total: average accuracy of total classification values. In italics is the performance of the best predictive model obtained among the eight tested.

Balanced datasets	$_{5min}SpO_2$ class	ANN			BACDT		
		Precision	Recall	F-score	Precision	Recall	F-score
Dataset 1	1	0.12	0.70	0.21	0.80	0.76	0.78
	2	0.16	0.43	0.23	0.61	0.56	0.59
	3	0.96	0.67	0.79	0.97	0.98	0.97
	Avg/total	0.88	0.65	0.73	0.94	0.94	0.94
Dataset 2	1	0.09	0.72	0.16	0.77	0.72	0.74
	2	0.09	0.47	0.16	0.57	0.53	0.55
	3	0.98	0.70	0.81	0.98	0.99	0.98
	Avg/total	0.93	0.69	0.78	0.96	0.97	0.97
Dataset 3	1	0.16	0.68	0.25	*0.80*	*0.76*	*0.78*
	2	0.26	0.42	0.33	*0.67*	*0.62*	*0.65*
	3	0.92	0.60	0.72	*0.95*	*0.96*	*0.96*
	Avg/total	0.80	0.58	0.65	*0.91*	*0.91*	*0.91*
Dataset 4	1	0.09	0.69	0.16	0.80	0.74	0.77
	2	0.12	0.47	0.19	0.58	0.54	0.56
	3	0.97	0.68	0.80	0.98	0.98	0.98
	Avg/total	0.92	0.67	0.76	0.96	0.96	0.96

For the artificial neural network, the variation of the number of hidden layers and number of neurons per hidden layer did not seem to have a significant effect on the model's classification performance (Table 3). As for the Bagged complex trees, the variation of the number of complex trees did not yield significant changes in classification performance (Table 4).

Table 3. Absence of impact on performance of the increase of neurons and hidden layers for artificial neural network (ANN). Example of the performance assessed by the F score on the balanced dataset 3 (see fig 2)

	ANN									
Hidden layers (n)		1			2			3		
Neurons/hidden layer (n)		10	50	100	10	50	100	10	50	100
F-score	$_{5min}SpO_2$ class 1	25	25	25	25	25	25	22	22	19
	$_{5min}SpO_2$ class 2	33	33	33	33	33	33	33	33	32
	$_{5min}SpO_2$ class 3	72	72	72	72	72	72	69	69	69

Table 4. Absence of impact on performance of the number of complex trees for bootstrap aggregation of complex decision trees (BACDT). Example of the performance assessed by the F score on the balanced dataset 3 (see Fig 2)

		BACDT	
		n = 30	n=50
F-score	$_{5min}SpO_2$ class 1	78	78
	$_{5min}SpO_2$ class 2	65	65
	$_{5min}SpO_2$ class 3	96	96

In agreement with previous studies regarding bagging being a better method for medical data classification, tree Bagging fared better than the artificial neural network used in this study [12]. It is noteworthy however that the gaps in performance results between the training and testing confusion matrices are relatively higher in the case of bagged trees model than in that of the artificial neural network (Fig 5). This seems to indicate that, although the bagged trees model was capable of learning very well from the data, there's still room for improvement in the generalization. The SMOTE algorithm is designed in such a way that should theoretically not affect the generalization of the trained model. In cases of extreme data imbalance, however, as is the case in this study, the over-sampling within the data space of a given minority class label, used for increasing the cardinality of the class label's set, is also likely to be extreme. This may render the data space of this class relatively dense with respect to the rest of the data, made up of real data points of the studied patient sub-population. This may potentially explain the classification model's relatively poor generalization for ${}_{5min}SpO_2$ class “1” and “2” with respect to the generalization for ${}_{5min}SpO_2$ class “3”. Also, since SMOTE generates synthetic data points by interpolating between existing minority class instances, it can obviously increase the risk of over-fitting when classifying minority class labels, since it may duplicate minority class instances. The fact that the training confusion matrix shows extremely high classification performances for the minority ${}_{5min}SpO_2$ class “1” and “2”, as opposed to those shown in the testing confusion matrix, suggests that the over-sampling of the minority ${}_{5min}SpO_2$ class using SMOTE could have caused some overfitting for these classes, but this would have to be further investigated.

Fig 5. Training and testing confusion matrices of artificial neural networks (ANN) and bootstrap aggregation of complex decision trees (BACDT) classifiers for SpO_2 prediction at 5 min following a ventilator setting change.

The strengths of this study include a large clinical database of mechanically ventilated children used with more than 7.10^5 rows. In a recent similar study in PICU, 200 patients were included with $1.15.10^3$ rows [15]. However, the volume of data is clearly insufficient. To use such machine learning predictive models, the pediatric intensive care community needs to combine multicenter high resolution database. In addition, children data could be pooled to neonatal and adult intensive care data, when possible, such as MIMIC III database [16]. The other strength is the process used to transform the data into a usable format and to correct a variety of artifacts present (S1 file). In health care, there is a significant interest in using clinical databases including dynamic and patient-specific information into clinical decision support algorithms. The ubiquitous monitoring of critical care units’ patients has generated a wealth of data which presents many opportunities in this domain. However, when developing algorithms domains, such as transport or finance, data are specifically collected for research purposes. This is not the case in healthcare where the primary objective of data collection systems is to document clinical activity, resulting in several issues to address in data collection, data validation and complex data analysis [17]. As detailed in S1 file, a significant amount of effort is needed, when data have been successfully archived and retrieved, to transform the data into a usable format for research.

This study has several limitations. The limited row number reduced the SpO_2 classification for machine learning predictive model to three clinically relevant classes. SpO_2 is a continuous variable and the use of three class is probably insufficient, especially when high SpO_2 range is suggested as potentially harmful [18, 19]. Instead of the classification model, the next step could be to test regression models' performance. SpO_2 was predicted at 5min after ventilator setting change, a clinically relevant delay. However, the delay between ventilator setting change and oxygenation steady state is not well defined and vary from 1 to 71 minutes according to the parameter set (FiO_2, PEEP or other parameters that change mean airway pressure) and clinical conditions studied [15, 20, 21]. This needs further research and probably more sophisticated clinical decision support systems using machine learning predictive models should consider these factors. Finally, we excluded hemodynamic unstable patients using a treatment criteria (≥ 2 vasoactive drugs infused) because this condition decreases pulse oximeter reliability [22, 23]. The validation and electronic availability of reliable markers of hemodynamic instability in children such as plethysmographic variability indices could be helpful [24].

Conclusion

This pilot study using machine learning predictive model resulted in an algorithm with good accuracy. To obtain a robust algorithm with such a method, more data rows are needed, suggesting the need of multicenter pediatric intensive care high resolution databases.

Acknowledgments

We would like to thank Mr. Redha Eltaani for his support in all tasks related to data access at Ste-Justine Hospital. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), by the Institut de Valorisation des Données (IVADO), by grants from the "Fonds de Recherche du Québec – Santé (FRQS)", the Quebec Ministry of Health and Sainte Justine Hospital.

References

1. Rose L, Schultz M, Cardwell C, Jouvet P, McAuley D, Blackwood B. Automated versus non-automated weaning for reducing the duration of mechanical ventilation for critically ill adults and children: a cochrane systematic review and meta-analysis. Crit Care. 2015;19:48. doi: 10.1186/s13054-015-0755-6.

2. Jouvet P, Eddington A, Payen V, Bordessoule A, Emeriaud G, Gasco R, et al. A pilot prospective study on closed loop controlled ventilation and oxygenation in ventilated children during the weaning phase. Crit Care. 2012;16(3):R85. doi: 10.1186/cc11343.

3. Flechelles O, Ho A, Hernert P, Emeriaud G, Zaglam N, Cheriet F, et al. Simulations for mechanical ventilation in children: review and future prospects. Crit Care Res Pract. 2013;2013:943281. doi: 10.1155/2013/943281.

4. Brossier D, El Taani R, Sauthier M, Roumeliotis N, Emeriaud G, Jouvet P. Creating a High-Frequency Electronic Database in the PICU: The Perpetual Patient. Pediatr Crit Care Med. 2018;19(4):e189-e98. doi: 10.1097/PCC.0000000000001460.

5. Cakar N, Tuŏrul M, Demirarslan A, Nahum A, Adams A, Akýncý O, et al. Time required for partial pressure of arterial oxygen equilibration during mechanical ventilation after a step change in fractional inspired oxygen concentration. Intens Care Med 2001;27(4):655-9.

6. Pediatric Acute Lung Injury Consensus Conference G. Pediatric acute respiratory distress syndrome: consensus recommendations from the Pediatric Acute Lung Injury Consensus Conference. Pediatr Crit Care Med. 2015;16(5):428-39. doi: 10.1097/PCC.0000000000000350.

7. Les recommendations des experts de la SRLF. Le monitorage et les alarmes ventilatoires des malades ventilés artificiellement. Réanim Urgences. 2000;9:407-12.

8. Chawla N, Japkowicz N, A. Kotcz A. Editorial: special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter. 2004;6:1-6.

9. Elhassan T, Aljurf M, Al-Mohanna F, Shoukri M. Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. Journal of Informatics and Data Mining. 2016;1:1-12.

10. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002;16:321-57. doi: 10.1613/jair.953.

11. Gnana Sheela K, Deepa S. Review on methods to fix number of hidden neurons in neural networks. Mathematical Problems in Engineering. 2013;2013:11. doi: 10.1155/2013/425740.425740.

12. Breiman L. Bagging predictors. Berkeley: University of California, Statistics Do; 1994 421.

13. Safavian S, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics. 1991;21(3):660-74. doi: 10.1109/21.97458.

14. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009;45(4):427-37. doi: 10.1016/j.ipm.2009.03.002.

15. Smallwood CD, Walsh BK, Arnold JH, Gouldstone A. Equilibration Time Required for Respiratory System Compliance and Oxygenation Response Following Changes in Positive End-Expiratory Pressure in Mechanically Ventilated Children. Crit Care Med. 2018;46(5):e375-e9. doi: 10.1097/CCM.0000000000003001.

16. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. doi: 10.1038/sdata.2016.35.

17. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine Learning and Decision Support in Critical Care. Proceedings of the IEEE Institute of Electrical and Electronics Engineers. 2016;104(2):444-66. doi: 10.1109/JPROC.2015.2501978.

18. Girardis M, Busani S, Damiani E, Donati A, Rinaldi L, Marudi A, et al. Effect of Conservative vs Conventional Oxygen Therapy on Mortality Among Patients in an Intensive Care Unit: The Oxygen-ICU Randomized Clinical Trial. JAMA. 2016;316(15):1583-9. doi: 10.1001/jama.2016.11993.

19. Pannu SR, Dziadzko MA, Gajic O. How Much Oxygen? Oxygen Titration Goals during Mechanical Ventilation. Am J Respir Crit Care Med. 2016;193(1):4-5. doi: 10.1164/rccm.201509-1810ED.

20. Tugrul S, Cakar N, Akinci O, Ozcan PE, Disci R, Esen F, et al. Time required for equilibration of arterial oxygen pressure after setting optimal positive end-expiratory pressure in acute respiratory distress syndrome. Crit Care Med. 2005;33(5):995-1000.

21. Fildissis G, Katostaras T, Moles A, Katsaros A, Myrianthefs P, Brokalaki H, et al. Oxygenation equilibration time after alteration of inspired oxygen in critically ill patients. Heart Lung. 2010;39(2):147-52. doi: 10.1016/j.hrtlng.2009.06.009.

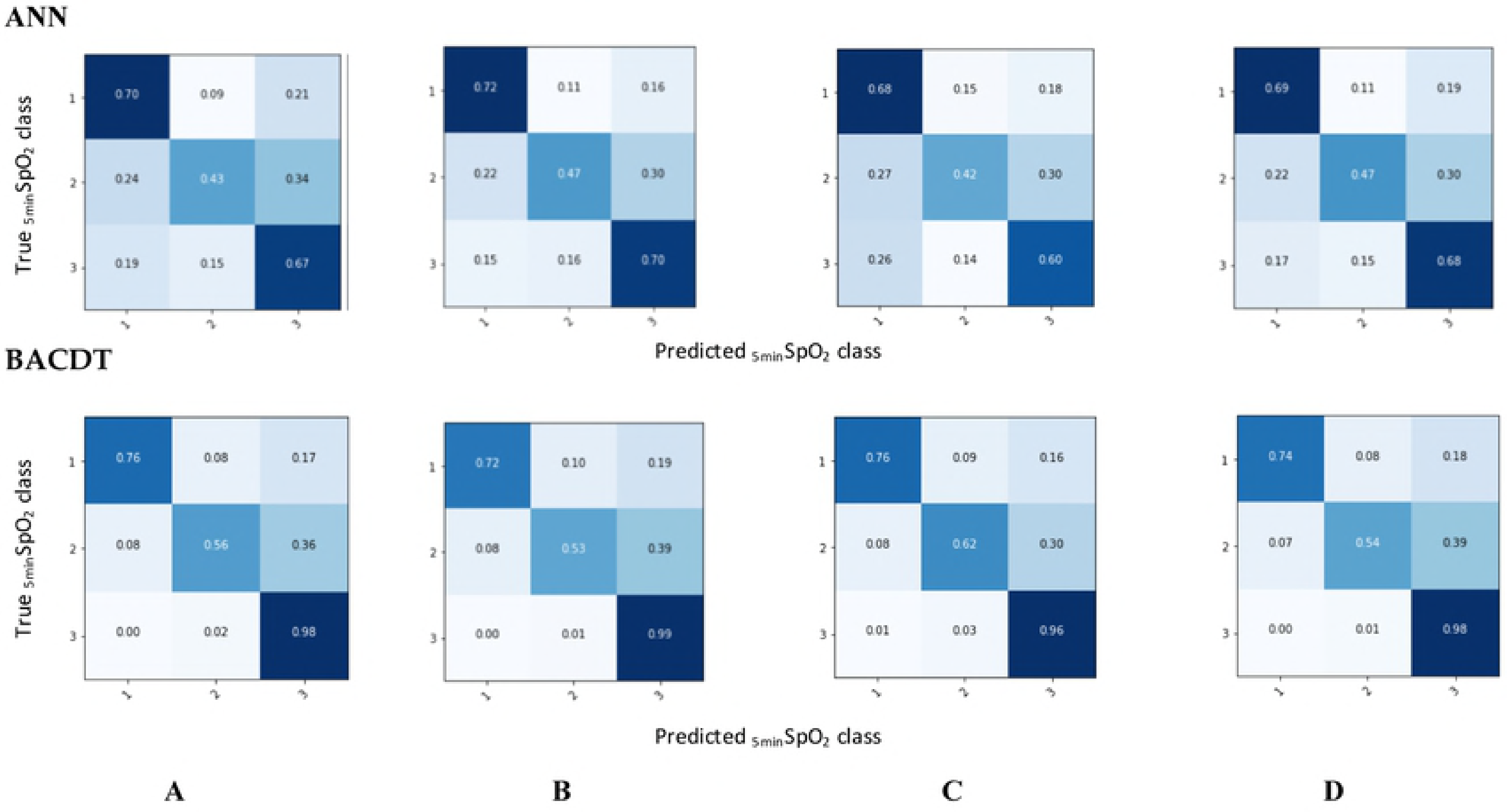
22. Salyer J. Neonatal and pediatric pulse oximetry. Respir care. 2003;48(4):386-96.

23. Fouzas S, Priftis KN, Anthracopoulos MB. Pulse oximetry in pediatric practice. Pediatrics. 2011;128(4):740-52. doi: 10.1542/peds.2011-0271.

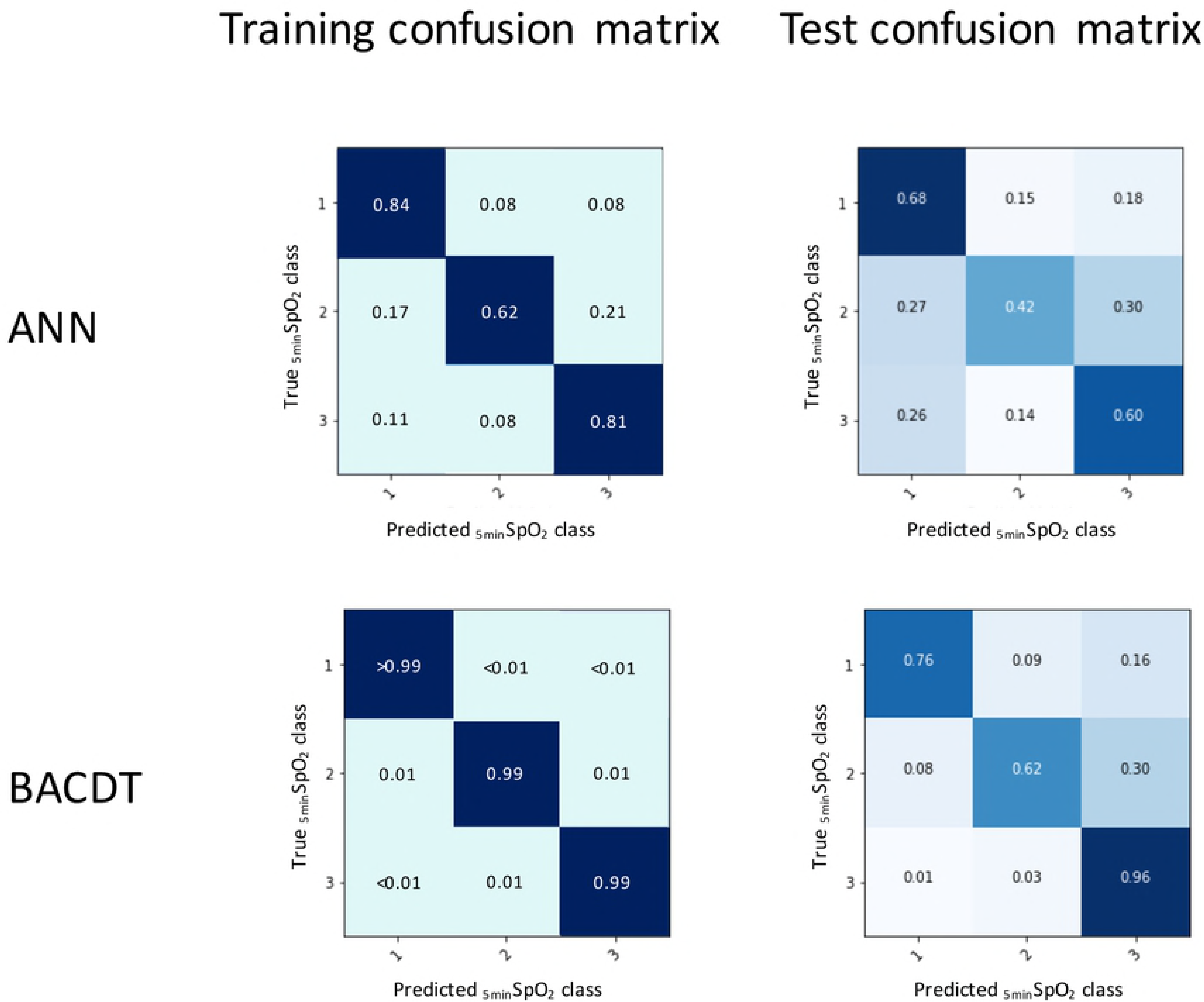
24. Chandler JR, Cooke E, Petersen C, Karlen W, Froese N, Lim J, et al. Pulse oximeter plethysmograph variation and its relationship to the arterial waveform in mechanically ventilated children. J Clin Monit Comput. 2012;26(3):145-51. doi: 10.1007/s10877-012-9347-z.

Supporting information

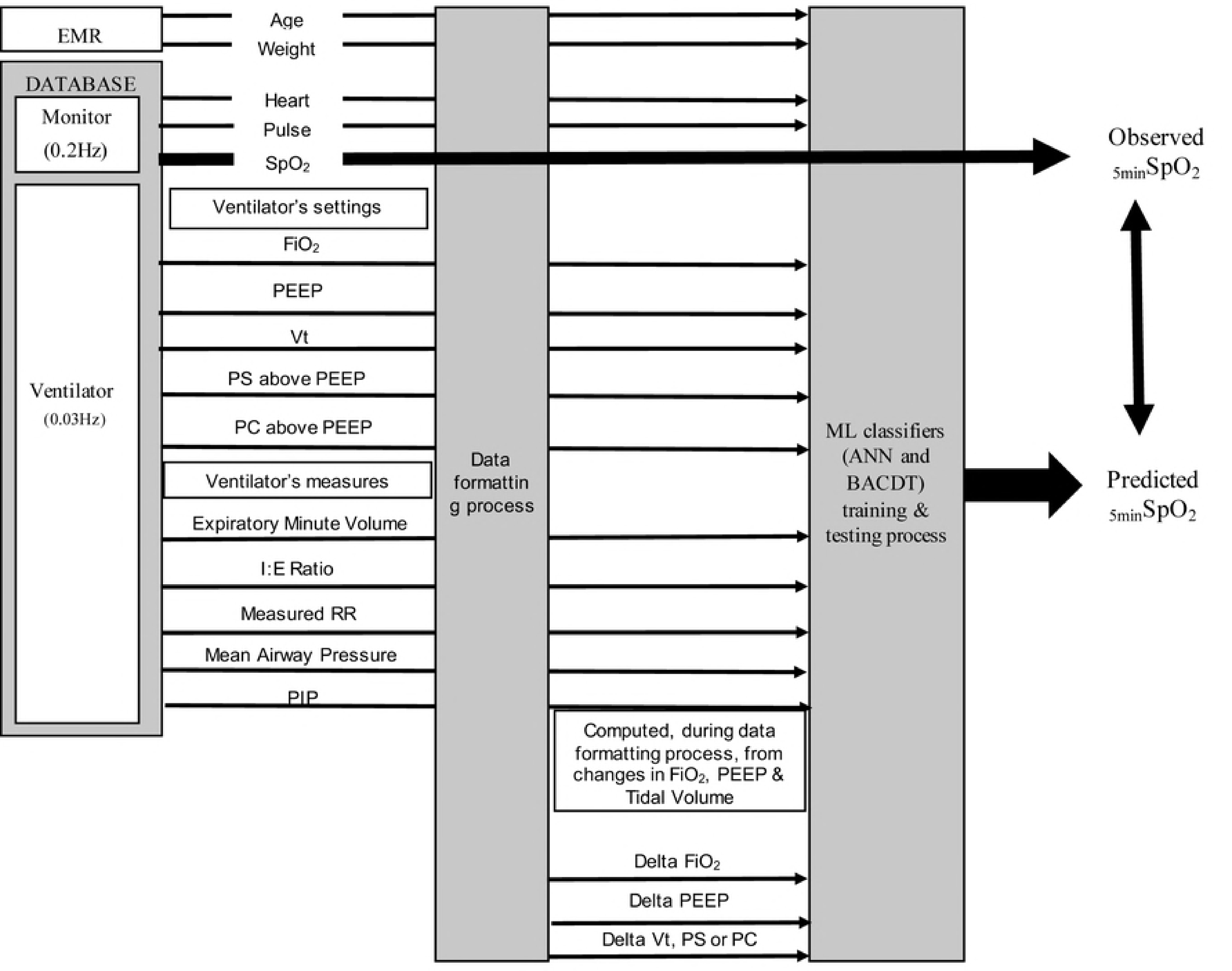
S1 File: Data formatting process



ANN
True $_{5min}SpO_2$ class
0.70 0.09 0.21
0.24 0.43 0.34
0.19 0.15 0.67
0.72 0.11 0.16
0.22 0.47 0.30
0.15 0.16 0.70
0.68 0.15 0.18
0.27 0.42 0.30
0.26 0.14 0.60
0.69 0.11 0.19
0.22 0.47 0.30
0.17 0.15 0.68
Predicted $_{5min}SpO_2$ class
BACDT
True $_{5min}SpO_2$ class
0.76 0.08 0.17
0.08 0.56 0.36
0.00 0.02 0.98
0.72 0.10 0.19
0.08 0.53 0.39
0.00 0.01 0.99
0.76 0.09 0.16
0.08 0.62 0.30
0.01 0.03 0.96
0.74 0.08 0.18
0.07 0.54 0.39
0.00 0.01 0.98
Predicted $_{5min}SpO_2$ class
A
B
C
D



Training confusion matrix
Test confusion matrix
ANN
BACDT
True 5minSpO2 class
Predicted 5minSpO2 class
1
2
3
0.84
0.08
0.08
0.17
0.62
0.21
0.11
0.08
0.81
0.68
0.15
0.18
0.27
0.42
0.30
0.26
0.14
0.60
>0.99
<0.01
<0.01
0.01
0.99
0.01
<0.01
0.01
0.99
0.76
0.09
0.16
0.08
0.62
0.30
0.01
0.03
0.96



EMR
Age
Weight
DATABASE
Monitor
(0.2Hz)
Heart
Pulse
SpO_2
Observed
$_{5min}SpO_2$
Ventilator's settings
FiO_2
PEEP
Vt
PS above PEEP
PC above PEEP
Ventilator
(0.03Hz)
Ventilator's measures
Expiratory Minute Volume
I:E Ratio
Measured RR
Mean Airway Pressure
PIP
Data formattin g process
ML classifiers (ANN and BACDT) training & testing process
Predicted
$_{5min}SpO_2$
Computed, during data formatting process, from changes in FiO_2, PEEP & Tidal Volume
Delta FiO_2
Delta PEEP
Delta Vt, PS or PC

Training set: 975,036 samples Test set: 193,528 samples Class Balancing: TOMEK applied to dataset (before dataset has been split into training & test set) to remove tomek links, random undersampling applied to class 3 once dataset is split into training and testing sub-sets, then SMOTE applied to classes 1 and 2 to make their cardinalities equal to that of class 3 (n=325,012).	Training set: 2,293,119 samples Test set: 201,926 samples Class Balancing: SMOTE applied to classes 1 & 2 to make their cardinalities equal to that of class 3 (n=764,373).	Training set: 487,464 samples Test set: 106,028 samples Class Balancing: TOMEK applied to dataset (before dataset has been split into training & test set) to remove tomek links, random undersampling applied to class 3 once dataset is split into training and testing sub-sets, then SMOTE applied to classes 1 and 2 to make their cardinalities equal to that of class 3 (n=162,488).	Training set: 1,462,503 samples Test set: 281,028 samples Class Balancing TOMEK applied to dataset (before dataset has been split into training & test set) to remove tomek links, random undersampling applied to class 3 once dataset is split into training and testing sub-sets, then SMOTE applied to classes 1 and 2 to make their cardinalities equal to that of class 3 (n=487,501).

- The data-set is first divided into two parts; the training-set and the test-set.
- The training of the "Bagged" Complex Trees includes a k-fold cross-validation, which is performed as follows:
  - ➢ Randomly partition the data-set into k equal-sized subsets (folds).
  - ➢ For each of the k equal-sized subsets:
  - ✓ Train/fit the model on the elements contained in the other (k-1) subsets.
  - ✓ Test the model's accuracy on the given subset.
  - ➢ Iterate over the k subsets, until each one has been used once for testing the model's performance during its training.
  - ➢ The training validation score consists of the average score obtained by validating the model on all k subsets.